Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37225

Multimodal Perception Based Recognition of Emotional Visual Narratives

Siping Zhang

College of Fine Arts and Calligraphy, Xihua University, Chengdu, China
wellington589125@gmail.com

Abstract. The recognition of emotional visual stories is an issue that involves simultaneous comprehension of semantic transformation between successive frames, the distribution of visual attention by the viewers, and observable emotional response. Information carried by visuals is not enough to completely characterize the perception process. For that reason, a trimodal computation model is designed with visual semantics, visual attentional behavior, and facial response. The visual stimuli consist of 120 four frame emotional stories. Sixty subjects watch the stories and identify the main emotion. Tobii Pro Fusion tracks eye movements with 120 Hz sampling frequency, whereas 30 fps video recording tool captures the facial videos simultaneously. CLIP ViT B 32 and Transformer compute the semantics of consecutive frames. BiLSTM computes the temporal representation of eye movements and facial action units individually. Next, cross-modal attention mechanism fuses the three types of representation together. With subject independent five fold cross validation, the trimodal model achieves an accuracy score of 0.824 ± 0.013 and macro F1 score of 0.817 ± 0.014, compared with scores 0.724 ± 0.017 and 0.706 ± 0.019 of visual only model. Fear and Surprise categories still have cross classification rate of 9% and 8%, respectively.

Keywords: multimodal perception, emotional visual narrative, emotion recognition, eye tracking, cross-modal attention

1. Introduction

Emotional visual narratives convey affect via interaction between cues, including scene semantics, behavior of characters, composition, and time, making emotion recognition a challenging perceptual task. Advances in deep visual models demonstrated an ability to learn affective representations in large image sets, proving that semantic features may be generalized better than manually designed color and texture features [1]. From the standpoint of temporal multimedia analysis, emotions are interpreted as dynamic interactions between scenes, actions, and facial cues, not as an attribute of an individual frame [2]. Thus, the current trend involves the combination of pretrained visual models and behavioral/physiological data. Synchronized modeling of narrative semantics, gaze and face, however, is rare despite the fact that gaze indicates selective attention, whereas facial behaviors indicate reactions to the same visual stimuli. A framework for multimodal perception is proposed for synchronization of three information channels, encoding of frame-level semantics with CLIP and temporal modeling, and fusion of modalities via cross-modal attention. Validation and ablation

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37225

experiments on subject-independent setting are conducted to analyze recognition accuracy, complementarity of modalities and generalization across observers in controlled emotional narrative scenarios.

2. Literature review

2.1. Evolution of visual emotion representation

The representation of visual emotions has moved from designed descriptors to deep learning-based representations of object/scene, facial expression, and semantic association. CLIP represents a further step in that direction since its image encoder can be employed to derive semantically aware frame representations based on image-text learning [3]. The recent work in the field of visual emotion analysis reveals that prompt-guided adaptation can enhance affective discrimination based on semantic connection between emotions and image representations [4]. While these innovations have helped increase semantic awareness, the task of recognizing narrative poses the challenge of the temporal dimension since emotions can develop over time by means of transformations over successive frames. Single-frame predictions might not catch transitions, causality, or relations between actors or events in the story.

2.2. Emotional cues in perceptual behavior

There is a possibility of getting evidence from perceptual behavior regarding how viewers direct their attention during interpretation of emotionally charged scenes. It has been noted through eye-tracking that the parts of images eliciting emotions direct attention and that the pattern of fixations varies depending on viewing time [5]. This means that it is possible to get cues for fixation duration, saccade length, pupillary difference, and transition between areas of interest as cues of priority allocation of elements of narratives. Facial behavior becomes another aspect to consider since action units become cues of reactions during the processing of stimulus [6]. The connection between gaze and facial activity depends on individuality and emotional uncertainty.

2.3. Cross-modal emotion modeling

Cross-modal emotion modeling aims at handling the representational heterogeneity problem and the temporal dependencies among those heterogeneous data. Research in the area of Multimodal Transformer indicates the ability of cross-modal attention to handle the interactions of sequence of different sampling and long dependency [7]. In addition, the emotion-recognition research indicates the benefits of using bidirectional attention to enhance the inter-modality mappings for semantically similar emotion classes [8]. The recent research on subspace mapping of representation also proves the necessity of considering both shared and non-shared features prior to fusion [9]. The above results justify the use of projection of visual, gaze, and facial features into the common latent space instead of directly concatenation. The remaining problem is how to align the temporal dependency between narrative events and perceptual responses of viewers.

DOI: 10.54254/2755-2721/2026.37225

3. Experimental methods

3.1. Synchronized multimodal experimental design

The visual stimuli are chosen from the publicly available Visual Storytelling Dataset, VIST. VIST was developed by Huang et al. in 2016. This first version of the dataset includes 81,743 images and 20,211 five-image sequences with the corresponding sequential image descriptions and stories. These resources provide an opportunity to choose visual stories which have a definite sequence of events. It should be noted that there is no information about the period of image acquisition in the original work. As a result, the year of the dataset is considered as the release year, which is 2016. Twenty sequences are chosen for each of six emotional states (joy, sadness, anger, fear, surprise, neutral). So, 120 stimuli are used totally. Four consecutive frames are chosen from each sequence. Sixty adults are asked to observe all sequences. Tobii Pro Fusion Eye Tracker captures eye positions, fixation duration, saccades, and pupil size dynamics at 120 Hz. At the same time, facial videos are recorded using a camera at 30 fps. After the observation of each sequence, participants determine the prevailing emotion.

3.2. Hierarchical perceptual feature encoding

The four visual frames are first processed using CLIP ViT B 32 to extract 512 dimensional semantic embeddings. Positional encoding is then added to preserve the temporal order of the four frames [10]. A Transformer Encoder models semantic relationships within the sequence, as shown in Equation 1. In this way, changes in characters, scenes, and events are represented within a unified visual narrative feature space.

$$H^{v}=\text{Transformer}\left([e_1^{v}+p_1,e_2^{v}+p_2,\ldots,e_T^{v}+p_T]\right) \tag{1}$$

Here, e_t^{v} denotes the CLIP feature of frame t, p_t denotes positional encoding, and T=4.

Gaze data are segmented using fixed time windows. Fixation duration, saccade amplitude, pupil change rate, and transition frequency between areas of interest are then extracted. OpenFace 2.0 is used to obtain facial action unit intensities from facial videos. The two continuous signals are separately processed using BiLSTM, as shown in Equation 2. This process captures temporal changes in viewing behavior and facial responses from both forward and backward directions.

$$h_t^{m}=\left[\overrightarrow{\text{LSTM}}\left(x_t^{m},h_{t-1}^{m}\right);\overleftarrow{\text{LSTM}}\left(x_t^{m},h_{t+1}^{m}\right)\right],\ m\in\{g,f\} \tag{2}$$

Here, g represents the gaze modality, f represents the facial modality, x_t^{m} denotes the input feature within a time window, and h h_t^{m} denotes the BiLSTM temporal representation.

3.3. Cross-modal attention fusion

Visual, gaze, and facial features are mapped into a unified 256 dimensional representation space through linear projection. They are then synchronized according to stimulus timestamps. Visual

DOI: 10.54254/2755-2721/2026.37225

features are used as the query matrix, while gaze and facial features form the key and value matrices [11]. Multihead cross modal attention calculates the correspondence between perceptual signals and visual narrative content, as shown in Equation 3.

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{3}$$

Here, Q, K and V denote the query, key, and value matrices, respectively. d_k denotes the dimension of the key vector.

The fused representation is passed through a fully connected layer and Softmax to obtain probabilities for the six emotional categories. Class weighted cross entropy is adopted during training to control differences in the number of valid samples across categories, as shown in Equation 4. AdamW is used for optimization together with an early stopping strategy and subject independent five fold cross validation. Data from the same participant are not included in both the training and test sets.

$$\mathcal{L}_{CE}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{6}w_c y_{ic}\log\hat{p}_{ic} \tag{4}$$

Here, N denotes the number of training samples, w_c denotes the class weight, y_{ic} denotes the ground truth label, and $\hat{p}_{ic}$ denotes the predicted probability for class c.

4. Results

4.1. Multimodal recognition performance

As illustrated in Table 1, there exists a clear trend of performance enhancement with the inclusion of further modalities. Specifically, the visual only approach attains a Macro F1 of 0.706 ± 0.019. When the gaze modality is included, the Macro F1 reaches 0.772 ± 0.016, which reveals the fact that gaze trajectories can offer the attention allocation information, which cannot be represented by visual semantics. The visual and facial combination obtains a score of 0.746 ± 0.017, which indicates a smaller gain compared to the previous combination of visual and gaze. As a result, facial responses offer supplementary classification information. The trimodal model can attain a Macro F1 of 0.817 ± 0.014 and an Accuracy of 0.824 ± 0.013. Both scores are greater than those for the visual only approach by 0.111 and 0.100, respectively. For the gaze and facial model, the Macro F1 score only achieves 0.678 ± 0.020. In the absence of visual semantics, it is impossible to reconstruct narrative context using viewing behaviors and responses alone.

Table 1. Comparison of emotion recognition performance across modality configurations

Model Configuration	Accuracy	Macro F1	Weighted F1
Visual	0.724 ± 0.017	0.706 ± 0.019	0.719 ± 0.018
Gaze	0.618 ± 0.022	0.601 ± 0.024	0.612 ± 0.023
Facial	0.592 ± 0.025	0.574 ± 0.026	0.586 ± 0.025

DOI: 10.54254/2755-2721/2026.37225

Table 1. (continued)

Visual + Gaze	0.783 ± 0.015	0.772 ± 0.016	0.779 ± 0.015
Visual + Facial	0.758 ± 0.016	0.746 ± 0.017	0.753 ± 0.016
Gaze + Facial	0.691 ± 0.019	0.678 ± 0.020	0.685 ± 0.019
Visual + Gaze + Facial	0.824 ± 0.013	0.817 ± 0.014	0.820 ± 0.014

4.2. Error pattern interpretation

As can be seen from Figure 1, correct recognition rates of the trimodal approach to simulate joy, sadness, anger, fear, surprise, and neutral are 88%, 82%, 80%, 78%, 82%, and 80% respectively. Performance variation among the six categories is reasonable. Confusion happens mostly between fear and surprise. Nine percent of the fear samples are categorized as surprise, while 8% of the surprise samples are recognized as fear. This suggests that although the two high arousal emotions are different now, they are still similar in suddenness and visual salience. Cross classification of about 8% also takes place between sadness and neutral, suggesting that the low arousal emotions are more dependent on context and viewing time. Attention weights should also be studied alongside these findings. In case gaze weights of facial regions and threatening objects are increased in fear samples, whereas corresponding facial action units are changing clearly, then the better performance results from visual cues, viewing paths, and observable reactions rather than only posterior probability.

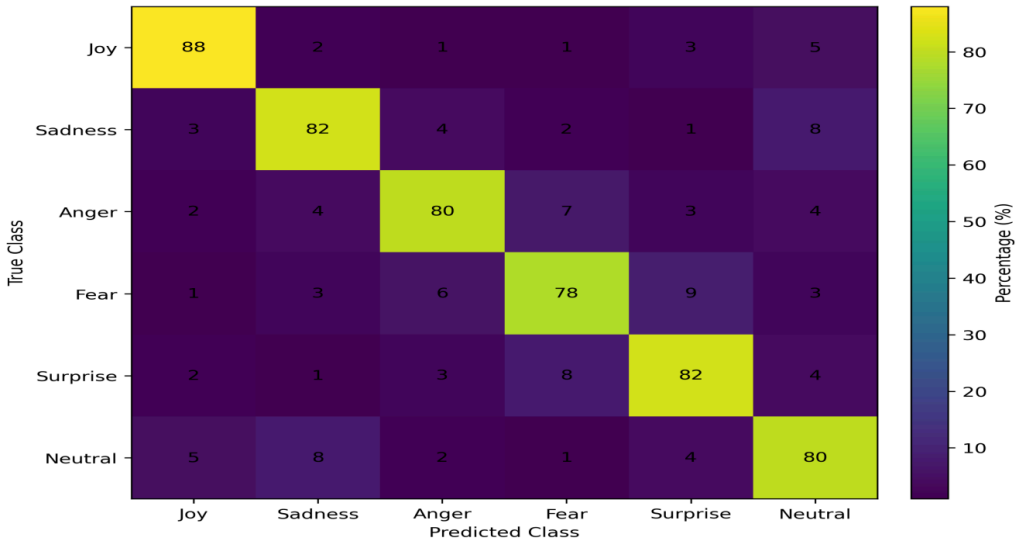


Figure 1. Normalized confusion matrix of six emotional categories for the trimodal model

5. Discussion

Trimodal fusion proves that gaze data is an essential additional feature for visual semantics. The configuration of Visual plus Gaze achieves a Macro F1 score of 0.772 ± 0.016 that is better than 0.746 ± 0.017 in case of the Visual plus Facial combination. It means that fixation location, duration, and transitions of gaze between areas of interest contain direct information regarding the selection of narrative cues. In turn, the Gaze plus Facial combination reaches the value of 0.678 ± 0.020 that indicates inability of behavioral reactions to restore the emotional context without visual semantics.

DOI: 10.54254/2755-2721/2026.37225

Fear and Surprise are the most often misclassified emotions due to similar salience and strong reaction to arousal. Trimodal fusion helps to avoid confusion of these two classes through integration of narrative cues, visual attention, and visible reactions. The cross-subjects stability of performance is dependent on viewing strategy and facial expressions of individuals.

6. Conclusion

A multimodal synchronous framework is developed for the identification of visual emotional narratives via visual stimulus, gaze activity, and facial response. CLIP and Transformer models are employed to extract the semantics and temporal relations from the four frame narratives using visual data. The BiLSTM model is used to encode the gaze trajectory and facial action unit sequences, while the cross modal attention model is used to model the interactions between the three modalities. Five fold validation on subject-independent data gives an Accuracy score of 0.824 ± 0.013 and a Macro F1 score of 0.817 ± 0.014 for the full trimodal model, surpassing the scores obtained by unimodal and bimodal models. The findings indicate the existence of complementary information between the visual semantics, attention behavior, and emotional response.

References

[1] Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion, 99*, 101847.
[2] Pan, B., et al. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing, 561*, 126866.
[3] Mocanu, B., Tapu, R., & Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing, 133*, 104676.
[4] Fu, B., et al. (2023). A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals. *Frontiers in Neuroscience, 17*, 1234162.
[5] Liu, S., et al. (2023). Multi-modal fusion network with complementarity and importance for emotion recognition. *Information Sciences, 619*, 679-694.
[6] Li, Y.-K., et al. (2023). MMFN: Emotion recognition by fusing touch gesture and facial expression information. *Expert Systems with Applications, 228*, 120469.
[7] Wu, C., et al. (2024). Towards visual emotion analysis via multi-perspective prompt learning with residual-enhanced adapter. *Knowledge-Based Systems, 295*, 111790.
[8] Fan, C., et al. (2024). Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion, 106*, 102306.
[9] Lu, N., et al. (2024). Bi-stream graph learning based multimodal fusion for emotion recognition in conversation. *Information Fusion, 106*, 102272.
[10] Pan, T., et al. (2024). Online multi-hypergraph fusion learning for cross-subject emotion recognition. *Information Fusion, 108*, 102338.
[11] Koupaei, M. R. E., & Daliri, M. R. (2024). DLEmotion: Deep learning-based emotion classification using visual attention location information. *Biomedical Signal Processing and Control, 95*, 106449.